Seventh Framework Programme
Theme 6 [SPACE]

Project: 607193 UERRA

Full project title:
Uncertainties in **E**nsembles of **R**egional **R**e-**A**nalyses

Deliverable D2.11
Probabilistic observation data set

WP no:	2
WP leader:	Met Office
Lead beneficiary for deliverable :	University of Bonn
Name of author/contributors:	Lilo Bach
Nature:	Report
Dissemination level:	PU
Deliverable month:	15
Submission date: March 27, 2015	Version nr: 1

A probabilistic observation data set for assimilation in ensemble nudging and statistical generation of upper-air pseudo observations

Lilo Bach[1], Christian Ohlwein[2], Andreas Hense[1] and Jan Keller[3]

[1]Meteorological Institute University of Bonn, [2]Hans-Ertel-Centre for Weather Research at the Meteorological Institute University of Bonn and [3]Deutscher Wetterdienst, Germany

Abstract

In this report, we present the work on deliverable D2.11. It contains an overview of the statistical methods developed for generating perturbed observations as well as pseudo observations, which is preparatory work for a high-resolution European ensemble reanalysis system. The work has been conducted by the Meteorological Institute of the University of Bonn (UB) as part of Work Package 2 on *Ensemble data assimilation regional reanalysis datasets* committed under the EU-FP7-funded collaborative project entitled *Uncertainties in Ensembles of Regional Reanalyses* (UERRA: Grant agreement no.: 607193, www.uerra.eu) in close cooperation with Christoph Schraff (data assimilation group at DWD) **D**eutscher **W**etter**d**ienst.
The report is divided into two parts. In the first part, the generation of an ensemble of reanalyses is outlined, whereby special emphasis is placed on the creation of perturbed observations following the purpose of the deliverable. In the second part, we describe the development of a statistical model for generating pseudo temperature observations that may be assimilated in areas or time-spans with sparse upper-air observations.

1 Probabilistic observations for the generation of a nudging ensemble reanalysis

UB's task as part of WP2 in UERRA is to provide a regional ensemble reanalysis system as well as a proof of concept high-resolution data set for Europe. In a first step, a hybrid LETKF/ensemble nudging system is developed which will be based on two DWD data assimilation systems:

- the nudging scheme (Schraff, 1997)
- the local ensemble transform Kalman filter for the convective scale (Reich, 2011).

A link of these two systems is considered particularly useful for reanalysis purposes as it combines their positive features yielding low RMSE (LETKF) and a smooth time

series with small error spikes (nudging) (Lei et. al, 2012a).
In the foregoing course of the project, the ensemble nudging component has been developed and tested, where the perturbed observations to be delivered in this report are an integral part of. Results on ensemble nudging will be presented in a later derivable. However, being conducted online as part of the ensemble nudging run, the description of the observation perturbation technique is inseparable from a short outline of ensemble nudging. In the following, we describe the development of ensemble nudging from deterministic nudging and place special emphasis on the generation of the probabilistic observations. Ensemble nudging is based on deterministic nudging, which is briefly overviewed in the next paragraph.

1.1 Deterministic Nudging

Nudging performs a continuous relaxation of the prognostic variables of any numerical weather prediction model towards observations during the forward integration of the model. Additional terms proportional to the observation-model equivalent departures are obtained by spatially spreading the observation increments to the target grid points. Thereby, a spatial weighting is performed using vertical and horizontal structure functions (Schraff, 1997). The temporal weighting function is designed such that observations are assimilated with maximal weight at the observation time. In contrast to intermittent 3-dimensional data assimilation schemes, asynoptic observations and high-frequency data can be assimilated at appropriate time. Nudging in its current implementation is not dependent on background or observation error covariance matrices. Instead, a static nudging coefficient having units of inverse time determines the strength by which the model state is corrected per model time step. Unlike 4d-Var or the ensemble Kalman filter, nudging the applied configuration of does not explicitly take into account flow-dependency. However, particularly due to its great performance-cost ratio yielding good analyses at low computational costs without dependence on tangent linear and adjoint models, nudging is used for many applications up to today (Stauffer and Seaman 1990, Stauffer et al. 1991, Seaman et al. 1995, Schraff 1997, Leidner et al. 2001, Otte et al. 2001, Deng et al. 2004, Deng and Stauffer 2006, Schroeder et al. 2006, Dixon et al. 2009, Ballabrera-Poy et al. 2009, Bollmeyer et al. 2015).

Due to the time-continuous manner in which the observations are assimilated, nudging yields smooth, physically consistent time series with little disturbance of the physical balances (e.g. Lei et al. 2012b). This is an advantage over intermittent techniques like the ensemble Kalman filter, where the sudden introduction of large numbers of observations often leads to strong error spikes in the assimilation time window (e.g. Hunt et al. 2004). Nudging is therefore considered an outstanding partner for techniques combining two different data assimilation schemes incorporating their respective advantages. Especially in reanalysis applications at high resolutions, a smoothness of time series should become an increasingly desirable feature for future developments.

1.2 Ensemble nudging using probabilistic observations

Applying ensemble nudging, the different ensemble members are nudged towards probabilistic observations. These are to be delivered in the framework of this report and shall be described in this paragraph.

Following e.g. Houtekamer et al. 1996, a probabilistic observation is given by perturbing the original observation *o* by means of a perturbation *o'* sampled from a normal distribution $o' \sim N(0, \sigma_o)$ with zero mean and a standard deviation given by the observation error σ_o. We assume normally distributed, unbiased, stationary in time as well as spatio-temporally uncorrelated observation errors. The latter is a wide-spread assumption often coming along with observation thinning and inflation of observation error variances (Lahoz et al. 2010).

We have implemented the perturbation process of observations into the limited area model COSMO as part of the nudging scheme. To provide physically sound observations, those exceeding reasonable value ranges are corrected accordingly. E.g., vertical lapse rates becoming super-adiabatic due to perturbation are corrected to prevent an extensive rejection of the probabilistic observations. In principal, observations from all used conventional observing systems including ACARS, AMDAR, TEMP, PILOT, WIND PROFILER, SYNOP, SHIP and DRIBU undergo the described perturbation process and a suitable quality control thereafter.

1.3 Specification of observation errors

The aim of ensemble nudging is to estimate the uncertainty of a nudging reanalysis given observation errors. Therefore it is of great importance to use reasonable estimates of the observation error standard deviations. However, the specification of observation errors and the corresponding covariances remain one of the major challenges in the field of data assimilation. An observation error as defined in this context depends both on the observation itself and the measurement process, but also on the data assimilation method and the resolution of the model it is assimilated into. It consists of a representativity component, a measurement component as well as a component resulting from uncertainty in the observation operators (Hollingsworth and Lönnberg, 1986).

Generally used methods for estimating observation errors are described in Hollingsworth and Lönnberg (1986) or Desroziers et al. (2005). They are diagnostics based on observations or on feedback output of data assimilation systems. Initially, we have investigated vertical and horizontal covariances of *o-b* and *o-a* departures (with *o* observation, *a* analysis and *b* background) from deterministic nudging reanalysis output (Bollmeyer et al. 2015). However, it has emerged that in the case of nudging the requirements for application of both Hollingsworth and Lönnberg (1986) or Desroziers et al. 2005 are not fulfilled:

- o-b cannot be computed straightforward for nudging. A quasi-observation independent first guess as given by least-squares methods does not exist for nudging as the model state is continuously corrected by data assimilation.

- *o-a* departures have little validity for nudging, because the static nudging coefficient prescribes how strong the model states will be pulled towards the observations. Therefore, *o-a* cannot give information about systematic errors.

After having figured out these limitations, we decided to rely on the observation error estimates used by DWD. These have been determined applying the aforementioned techniques of Hollingsworth and Lönnberg (1986) and Desroziers et al. (2005) to feedback data from other non-convection resolving NWP systems of similar resolution like COSMO. The latter is of particular importance to guarantee for a reasonable estimation of the representativity component. The DWD observation error estimates have mainly been used for the quality control in the regional NWP system. Recently, their magnitude has been rechecked and partly reconfirmed or updated using feedback output from the new LETKF data assimilation scheme.

Assimilated using ensemble nudging, the probabilistic observations have yielded reasonable ensemble spread and verification results in different case studies. The results are part of a deliverable on the feasibility of the hybrid LETKF-ensemble nudging system to be developed and will therefore be shown at a later stage of the project.

2 Pseudo observations for observation-sparse areas and time spans

In data assimilation, vertical correlations between observations expressed as structure functions play an important role for spreading observational information to the target grid points. Here, we make use of vertical correlations to statistically derive pseudo upper-air temperature observations from near surface temperature observations. These may help to enhance the data coverage in areas or time spans sparse of upper-air observations.

2.1 Statistical model

If vertical temperature profiles are to be derived, canonical correlations between 2m temperature and radiosonde data can be utilized. These are advantageous as they also take into account the correlations between temperatures at different radiosonde measurement levels. To examine the feasibility of our approach, we derive pseudo temperatures for just one atmospheric level. For that purpose, we perform a multiple linear regression analysis between German T_{2m} data and Era-Interim 850-hPa data.

SYNOP station data in an area of influence of 100 km are chosen for each Era-Interim grid point for a period ranging from 1979 to 2008. Only stations available for the whole time span are taken into account. Possible gaps in the data are filled using an expectation-maximization algorithm. For the regression, the data are sub-divided according to months and time and centered.

We obtain *n* regression coefficients (where *n* is the number of T_{2m} observations corresponding to the grid point) which remain stable when cross-validation is applied. Based on these coefficients plus real-time 2m temperature data, pseudo temperatures can be derived and assimilated. It is of great advantage that the model can be utilized at any time - under the assumption of a constant statistical relationship, respectively.

2.2 Results

The resulting multiple correlations are best for the summer months around noon. Here, correlations of more than $r^2=0.9$ are obtained in a wide area of Germany. This can be expected, because at this time the boundary layer is mostly convective and of considerable height. Similar heights can be reached under westerly atmospheric conditions that come along with strong advection and thus a neutral boundary layer. Lower multiple correlations are expectable for boundary layers of low height. These occur under stable conditions leading to a physical decoupling, i.e. a deterioration of the linear relation between the 2m temperature and the one in 850 hPa. Stable boundary layers develop for example in summer nights after sunny days when the radiation budget gets negative. This is often accompanied by the development of a residual layer. Moreover, stable boundary layers develop during winterly high-pressure systems over Central Europe (Stull, 1988). Indeed, lower multiple correlation coefficients are found for the winter months and summer nights.

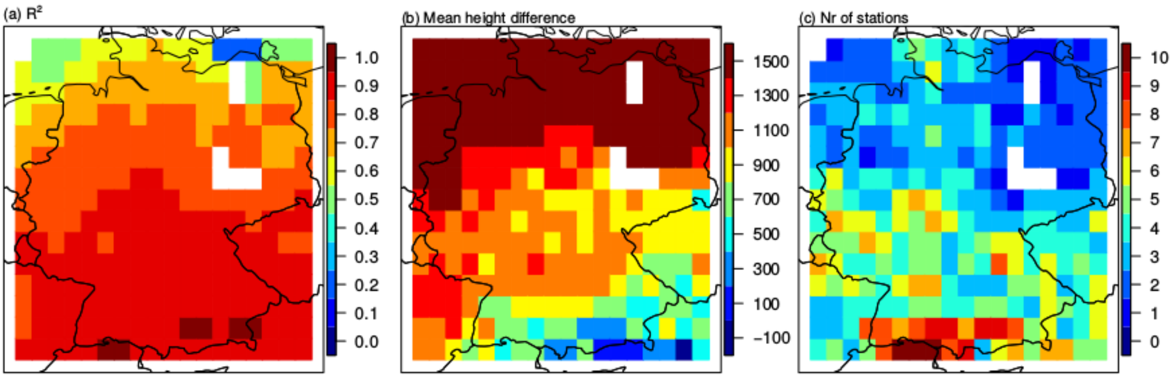

Figure 1: Results from multiple linear regression for Germany, June 12 UTC. (a) squared multiple correlation coefficient, (b) mean difference between 1500 m and mean T2m station height, (c) number of stations with a pvalue lower than 0.05.

Figure 1 shows exemplarily the regression results for June, 12 UTC. Figure 1(a) depicts the squared multiple correlation coefficient. Apparently, it increases from a value of 0.5 (the square root is still approximately 0.7) at the German coasts of North and Baltic Sea coasts to 0.9 in the southern part of Germany. A comparison with Figure 1(b) showing the height difference between the climatological geometrical height of 850 hPa, i.e. 1500 m, and the average height of the SYNOP stations measuring T_{2m} it becomes obvious that the correlation increases with decreasing height difference. As shown by Figure 1(c) there is also a connection to the number of stations contributing significantly to the regression (having a pvalue less than 0.05) - albeit a fairly weaker one. Generally, in the coastal areas and eastern Germany, noticeably less SYNOP stations are available. The white areas in the plots represent grid points whose surroundings lack continuously recorded SYNOP data.

As outlined, the developed statistical model for deriving pseudo temperatures at the 850 hPa pressure level from 2m-temperatures yields high multiple correlations. To better represent the actual lapse rate, humidity can be utilized as a further covariate. In the coastal areas additional use of sea surface temperatures can be taken into consideration to enhance the multiple correlation. Vertical profiles for the lower troposphere can be derived employing canonical correlation analysis. For long-time reanalyses such as 20-CR, the regression coefficients can be generated flow-dependent using a Kalman-MOS. Currently, extensive assimilation experiments are run to achieve an estimate of the analysis quality.

References

Ballabrera-Poy, J., Kalnay, E., and Yang, S. (2009). Data assimilation in a system with two scales – combining two initialization techniques. *Tellus*, 61A: 539-549.

Bollmeyer, C., Keller, J., Ohlwein, C., Wahl, S., Crewell, S., Friederichs, P., Hense, A., Keune, J., Kneifel, S., Pscheidt, I., Redl, S., and Steinke, S. (2015). Towards a high-resolution reanalysis for the European CORDEX domain. *Quaterly Journal of the Royal Meteorological Society*, 141:1-15.

Deng, A., Seaman, N., Hunter, G., and Stauffer, D. (2004). Evaluation of interregional transport using the MM5-SCIPUFF system. *Journal of Applied Meteorology*. 43:1864-1886.

Deng, A. And Stauffer, D. (2006). On improving 4-km mesoscale model simulations. *Journal of Applied Meteorology*, 45:361-381.

Desroziers, G., Berre, L., Chapnik, B., and Poli, P. (2005). Diagnostics of observation, background and analysis-error statistics in observation space. *Quaterly Journal of the Royal Meteorological Society*, 131:3385-3396.

Dixon, M., Li, Z., Lean, H., Roberts, N., and Ballard, S. (2009). Impact of data assimilation on forecasting convection over the Unitied Kingdom using a high-resolution version of the Met Office unified model. *Monthly Weather Review*, 137:1562-1585.

Hollingsworth, A. An Lönnberg, P. (1986). The statistical structure of short-range forecast errors as determined from radiosonde data. Part 1: The wind field. *Tellus*, 38A:111-136.

Houtekamer, P., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H. (1996). A system simulation approach to ensemble prediction. *Monthly Weather Review*, 124:1225-1242.

Hunt, B.R., Kalnay, E., Kostelich, E., Ott, E., Patil, D., and Co-Authors. (2004). Four-dimensional ensemble Kalman filtering. *Tellus*, 56A:273-277.

Lahoz, W., Khattatov, B., and Ménard, R., editors. (2010). *Data Assimilation. Making Sense of Observations.* Springer.

Lei, L., Stauffer, D., and Deng, A. (2012b). A hybrid nudging-ensemble Kalman filter approach to data assimilation. Part II: Application in a shallow-water model. *Tellus*, 64.

Lei, L., Stauffer, D., Haupt, S.E., and Young, G. (2012a). A hybrid nudging-ensemble Kalman filter approach to data assimilation. Part I: Application in the Lorenz system. *Tellus A*, 64.

Leidner, S., Stauffer, D., and Seaman, N. (2001). Improving short-term numerical weather prediction in the California coastal zone by dynamic initialization in the marine boundary layer. *Monthly Weather Review*, 129: 275-294.
Otte, T., Seaman, N., and Stauffer, D. (2001). A heuristic study on the importance of anisotropic error distribution in data assimilation. *Journal of Applied Meteorology*. 47:1853-1867.
Reich, H., Rhodin, A., and Schraff, C. (2011). LETKF for the nonhydrostatic regional model COSMO-DE. Technical report, Deutscher Wetterdienst, Offenbach, Germany.
Schraff, C. (1997). Mesoscale data assimilation and prediction of low stratus in the Alpine region. *Meteorology and Atmospheric Physics,* 64:21-50.
Schraff, C. And Hess, R. (2012). A description of the nonhydrostatic regional COSMO-model. Part III: Data assimilation. Technical report, Deutscher Wetterdienst, Offenbach, Germany.
Schroeder, A., Stauffer, N., Seaman, N., Deng, N., Gibbs, A., and Co-Authors (2006). An automated high-resolution, rapidly relocatable meteorological nowcasting and prediction system. *Monthly Weather Review*, 134:1237-1265.
Seaman, N., Stauffer, D., and Lario-Gribbs, A. (1995). A multi-scale four-dimensional data assimilation system aoolied in the San Joaquin Valley during SARMAP. Part I: Model design and basic performance characteristics. *Journal of Applied Meteorology*. 34:1739-1761.
Stauffer, D. and Seaman, N. (1990). Use of four-dimensional data assimilation in a limited-area mesoscale model. Part I: Experiments with synoptic data. *Monthly Weather Review*, 118:1250-1277.
Stauffer, D., Seaman, N., and Binkowski, F. (1991). Use of four-dimensional data assimilation in a limited-area mesoscale model. Part II: Effects of data assimilation within the planetary boundary layer. *Monthly Weather Review.* 119:734-754.
Stull, R. (1988). *An Introduction to Boundary Layer Meteorology*. Kluwer Academic Publishers.